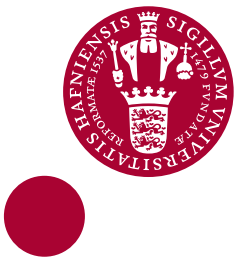

Measuring statistical anxiety and attitudes toward statistics: The development of a comprehensive Danish instrument (HFS-R)

Nielsen, Tine; Kreiner, Svend

Published in:
Cogent Education

DOI:
10.1080/2331186X.2018.1521574

Publication date:
2018

Document version
Publisher's PDF, also known as Version of record

Nielsen & Kreiner, *Cogent Education* (2018), 5: 1521574
https://doi.org/10.1080/2331186X.2018.1521574

cogent education

EDUCATIONAL ASSESSMENT & EVALUATION | RESEARCH ARTICLE

Measuring statistical anxiety and attitudes toward statistics: The development of a comprehensive Danish instrument (HFS-R)

Tine Nielsen[1]* and Svend Kreiner[2]

Received: 12 July 2018
Accepted: 05 September 2018
First Published: 10 September 2018

*Corresponding author: Tine Nielsen, Department of Psychology, University of Copenhagen, Copenhagen, Denmark
E-mail: tine.nielsen@psy.ku.dk

Reviewing editor:
Sammy King Fai HUI, The Education University of Hong Kong, Hong Kong

Additional information is available at the end of the article

Abstract: Motivated by experience with students' psychological barriers to learning statistics, we modified and extended the Statistical Anxiety Rating Scale (STARS) to develop a contemporary and valid (face, content, criterion and construct) Danish measure of attitudes and relationship towards statistics for use with higher education students taking statistics within another discipline. Two subscales were excluded because of lack of conceptual unidimensionality or derogatory content, and single items were modified for face and content validity enhancement in the remaining subscales. Following a pilot study and main study, the resulting 26-item Danish instrument (HFS-R for "holdninger og forhold til statistik - Revideret", in English "Attitudes and Relationship to Statistics - Revised") consists of four subscales: Test and Class Anxiety (TCA), Interpretation Anxiety (IA), Fear of Asking for Help (FAH), and Worth of Statistics (WS). Each scale was analyzed using Rasch and graphical log-linear Rasch models. The FAH subscale fits the Rasch model, whereas the TCA, IA, and WS subscales each fit graphical log-linear Rasch models (GLLRMs) each with evidence of differential item functioning (DIF). One TCA item functioned differentially relative to age, one WS item relative to statistics course (first or second), and two IA items relative to statistics course and academic discipline

ABOUT THE AUTHORS

Tine Nielsen is an Associate Professor at the Department of Psychology, University of Copenhagen, Denmark. Her research is focused within the fields of higher education psychology, health psychology, and psychometrics, under the heading PsychMeasure. The present study is one in a series of validity studies aimed at providing the Danish educational research community with valid and reliable measures. She is an active member of several education as well as psychometric societies and network groups, e.g. the European Association for Research in Learning and Instruction (EARLI), The European Rasch Research and Teaching Groups (ERRTG), and the Danish Measurement network (DM), which she founded in 2016. She is a subject editor (higher education and psychometrics) at the Scandinavian Journal of Educational Research.

Svend Kreiner is a Professor Emeritus at the Biostatistics Unit at the Department of Public health, University of Copenhagen.

PUBLIC INTEREST STATEMENT

The research reports the development and the validity of the Danish contemporary instrument (HFS-R) for the measurement of attitudes and relationship to statistics for use in higher education research. Analyses were performed within the framework of item response theory (IRT) by using Rasch family measurement models. These models where chosen as they set strict standards for measurement quality. The results show that the four short subscales (Test and Class Anxiety, Interpretation Anxiety, Fear of Asking for Help, and Worth of Statistics) have good psychometric properties, although three of the scales were in need of adjustment for differential item functioning to measure fairly across subgroups of students. The three anxiety scales do not make up a single unidimensional scale, as suggested by other research. The scales are of sufficient quality to warrant their use in contemporary higher education research.

(sociology, public health). The IA and TCA subscales were well targeted to the study population, while targeting of FAH and WS was poorer. Unidimensionality across the three anxiety subscales (FAH, TCA, and IA) was tested and clearly rejected. The HFS-R was found to be of sufficient psychometric quality to warrant its use in higher education research and teaching. We recommend that unidimensionality should be formally tested before using composite scores across anxiety subscales in the original STARS.

Subjects: Mathematics & Statistics; Psychological Methods & Statistics; Testing, Measurement and Assessment; Educational Psychology

Keywords: statistical anxiety; attitudes toward statistics; construct validity; Rasch analysis; higher education

1. Introduction

The present study arose from a long interest in the psychological barriers against learning statistics, when statistics is a tool subject within another academic discipline, as we had experienced personally such barriers through many years of teaching. Thus, we wanted to explore this phenomenon with Danish university students, and on the basis of the results, be able to work with course designs and approaches to teaching aimed at decreasing students' statistical anxiety where necessary. We therefore aimed to translate, adapt, or develop a valid instrument for this purpose.

Psychological barriers to learning statistics have been researched as occurrences of statistical anxiety and attitudes toward statistics through more than six decades. Previous research has found that attitudes toward statistics and statistical anxiety are associated with student performance in statistics classes. Thus, as early as 1954, Bendig and Hughes found that psychology students' attitudes toward statistics could account for just 5% of the variation in their final grades. For graduate students from non-mathematically oriented disciplines, statistics, Onwuegbuzie and Seaman (1995) found test anxiety to correlate negatively with students' final exam grades. Galli and colleagues found that the mean anxiety score for psychology students failing the final statistics exam was significantly higher than that of students passing the exam (Galli, Ciancaleoni, Chiesi, & Primi, 2008). Statistical anxiety and attitudes toward statistics have also more recently been shown to be related to strategies of learning (Kesici, Baloğlu, & Deniz, 2011) and inversely related to attitudes toward science (Bui & Alfaro, 2011). The field literature includes many recommendations on how to teach in order to relieve or avoid students' statistical anxiety. However, we have only been able to locate a single study directly showing the effect of specific teaching approaches or course design on students' statistical anxiety—Williams' (2010) study on statistics anxiety and instructor immediacy (i.e. communicate at a close distance, smile, engage in eye contact, use direct body orientation, etc.). Using a pretest-posttest-control group design, she found that instructor immediacy was significantly associated with the six dimensions of statistical anxiety measured with the STARS (see below), and that immediacy explained between 6% and 20% of the variance in these dimensions. Thus, much remains to be studied in this regard.

Based on previous research, Chew and Dillon (2014) argue that the distinction between attitudes toward statistics and statistical anxiety provides more detailed insights than when these constructs are not sufficiently separated. Furthermore, it has been established that statistical anxiety is related to, and perhaps even somewhat overlapping with, albeit not the same as, mathematical anxiety, and it should thus be studied in its own right (see Chew & Dillon, 2014 for details).

An initial literature search for an appropriate measure revealed that a much-used instrument for this purpose was the Statistical Anxiety Rating Scale (STARS; Cruise & Wilkins, 1980). Chew and Dillon (2014), in their review, reported discovering five named instruments for measuring statistics

anxiety: the Statistics Anxiety measure (Earp, 2007), the Statistical Anxiety Scale (Virgil-Colet, Lorenzo-Seva & Condon, 2008), the Statistics Anxiety Inventory (Zeidner, 1991), the Statistics Anxiety Scale (Pretorius & Norman, 1992), and the STARS (Cruise & Wilkins, 1980). Of these, only the STARS and the Statistics Anxiety Measure contained both anxiety and attitude scales, and the STARS was found to be by far the most widely used, namely by 78% of the 50 articles reviewed. We therefore decided to base our work on the STARS.

Several lines of research have looked into the validity of the STARS and the way it is scored. Thus, recently, based on a non-fitting model (confirmatory factor analysis) with a very small sample, Hsiao (2010) suggested that the STARS is two-dimensional, rather than having six separate dimensions, and that composite scores should be formed across the anxiety and attitude subscales, separately. Hsiao's proposal has since been explored further using confirmatory factor analysis in larger samples by a group of German researchers, who also came to the conclusion that the STARS is two-dimensional (Macher, Paechter, Papousek, & Ruggeri, 2011; Macher et al., 2013; Papousek et al., 2012). On this basis, they also proposed that composite scores should be formed across the anxiety and attitude subscales, and used in place of the six subscales as originally proposed (Cruise, Cash, & Bolton, 1985). However, a most recent confirmatory factor analysis study (DeVaney, 2016) found support for the original six-factor structure and that the anxiety and attitude factors were correlated.

A number of studies have focused on group-wise differences in statistical anxiety and attitudes toward statistics, as measured by the STARS subscales. For example, a study by Baloğlu, Deniz, and Kesici (2011) examined differences on the STARS for groups defined by country, gender, grade point average, and age and found country-wise differences on four of the STARS subscales, as well as gender differences on two subscales. Similarly, Hsiao and Chiang (2011) found gender differences with the STARS for a small Taiwanese sample. However, none of these studies have ascertained whether the STARS subscales are indeed measurement invariant for the groups compared, i.e. free of differential item functioning (DIF), and thus that these differences are true and not spurious findings.

To our knowledge, only two previous studies (Maat & Rosli, 2016; Teman, 2013) have employed the item response theory (IRT) to study the validity of the STARS. Maat and Rosli (2016) conducted a Rasch analysis of a Malay version of the STARS with a small sample of postgraduate students in different disciplines, including mathematics education. However, it appears that the analysis was conducted on a total scale comprising all STARS items and that no DIF analyses were conducted in the study. Teman (2013) conducted Rasch analysis for each of the six STARS subscales with a sample of 431 American university students. Teman specifically used the so-called "rating scale model" (Andrich, 1978), which is simply an ordinal Rasch model (or partial credit model; PCM), where item thresholds are constrained to be equal across all items in a scale. Teman's (2013) analysis resulted in as many as 20 of the 51 items being excluded as they did not fit the subscale models. Teman's (2013) analysis included DIF analysis relative to gender and study level (undergraduate versus graduate) and found no evidence of gender DIF, while evidence of DIF relative to study level was found for six items. Overall, Teman (2013) concluded that only 21 items across the four subscales comparable to the scales employed in the present study were valid, and among these, three items should be adjusted for DIF relative to the study level. Across the entire six STARS scales, just 31 items were found to be valid, and six of these suffered from DIF. Teman (2013) suggested that future research on the STARS, employing Rasch analysis, should be conducted with larger samples, across a wider range of academic majors, and both within and between several universities, as this would allow for more advanced DIF analyses across these subgroups.

1.1. The current study

An initial examination of the subscales and the items of the STARS revealed that both language translation and some degree of adaptation were necessary for our purposes (i.e. adaptations needed to ensure better face and content validity), as well as consideration of the two parts,

namely statistical anxiety and attitudes (c.f. the recent suggestion of composite scores). Taking the STARS as a starting point, the aim of the current study was thus to develop a contemporary Danish measure of statistical anxiety and attitudes toward statistics, of good psychometric quality for use with higher education students taking statistics as a tool subject, using IRT methods. This included assessment of criterion-related construct validity as defined by Rosenbaum (1989) and assessment of criterion validity by analysis of the correlational relationships between subscales of the instruments for students from different academic disciplines and students taking their first and second statistics course, respectively.

2. Methods

2.1. Instrument

The STARS (Cruise & Wilkins, 1980) is a self-report questionnaire intended to measure statistical anxiety and attitudes toward statistics, each by three subscales. It has been translated into several languages (e.g. German, Chinese, and Turkish), but not hitherto into Danish. The STARS consists of 51 items making up the six subscales: Test and Class Anxiety (TCA) with eight items, Interpretation Anxiety (IA) with 11 items, Fear of Asking for Help (FAH) with four items, Worth of Statistics (WS) with 16 items, Computation Self-Concept (CSC) with seven items, and Fear of Statistics Teachers (FST) with five items. For the three anxiety subscales—TCA, FAH, and IA—students use a five-point response scale, 1 (none) to 5 (high), to indicate the level of anxiety they would experience in the situations given in the items. For the three attitude subscales—WS, SCS, and FST—students use a five-point response scale, from 1 (strongly disagree) to 5 (strongly agree) to indicate the extent to which the statements describe their feelings or attitudes. The response scales only have meaning anchors for the first and last categories, and thus have only numbers for the three middle categories. All item statements are related to the field of statistics in some way or another. Item statements for the attitude subscales are mixed in between each other in a set order, as are the items for the anxiety subscales.

2.2. Translation and piloting

Initial translation and adaptation of the STARS were done in a series of steps. First, it was translated into Danish by a trained translator. Second, two independent evaluations of the translated items of each subscale were made by expert judges (psychology/psychometrics and statistics/psychometrics) with regard to face and content validity as well as both time-wise and tone-wise appropriateness, given that the STARS was developed in the late 1970s and that some items express a derogatory tone toward statisticians. Third, the two evaluations were compared and a consensus reached concerning the adaptation and development needs of the STARS.

The consensus decision was to exclude two subscales from the instrument. The CSC subscale was excluded, as it appeared not to be conceptually unidimensional (i.e. it included mathematical problems/anxiety and statistics anxiety in a mixture, e.g. *Since I've never enjoyed math, I don't see how I can enjoy statistics* and *Statistics isn't really bad. It's just too mathematical*). The FST subscale was excluded as both the title and items appeared inappropriate for a researcher to ask a statistics teacher to ask his/her students due to their derogatory content toward statisticians (e.g. *Statistics teachers are so abstract they seem inhuman*and *Most statistics teachers are not human*). In the remaining four subscales, further decisions were made as to whether single items should be included, adapted, or excluded, in order to improve face and content validity:

- TCA: five original items were included. The last three items referred to two phenomena that are not commonplace in Danish higher education: testing students during courses and enrolling in single courses as Danish students usually follow entire degree programs. These items were all replaced with new items referring to statistics classes and taking statistics exams.

- FAH: three original items were included. One further item appeared outdated as it referred to *asking someone in the computer center for help in understanding a printout*. The item was therefore modified to reflect asking for help in exercise classes. Finally, a fifth item was added to cover a missing area in the scale, namely asking for further elaboration of something in statistics lectures.
- IA: five original items were included. Three further items were modified, as they referred to research projects and abstracts from journal articles, which are not commonly a part of statistics classes in Danish higher education. The items were modified to reflect instead exercises and course tasks. The last three items were excluded; two, because they referred to uncommon or outdated activities such as reading automobile advertisements and assessing odds in a lottery, while the last was excluded because it did not address any aspect of IA (i.e. *arranging to have a body of data put into the computer*).
- WS: one original item was included, while seven items were modified to make them simpler by avoiding double statements or to make them neutrally or positively worded, as the scale contained some very negatively phrased items (e.g. *I don't understand why someone in my field needs statistics* was modified to the positive form *Statistics is useful*). Furthermore, eight items were excluded as they did not add to the content of the scale and were therefore considered content-wise redundant.

The response scales were modified by reducing their number of categories from five to four and by adding appropriate meaning anchors for all categories. Lastly, the instructional statements were modified to fit the response scales.

The resulting Danish HFS instrument ("Holdning og Forhold til Statistik", in English "attitudes towards and relationship with statistics") consisted of 29 items divided on four scales: TCA (8), FAH (5), IA (8), and WS (8), with one four-point response scale (1 = no anxiety, 2 = a little anxiety, 3 = some anxiety, and 4 = a lot of anxiety) for the three anxiety subscales (TCA, FAH, and IA) and another four-point response scale (1 = definitely disagree, 2 = disagree more than agree, 3 = agree more than disagree, and 4 = definitely agree) for the attitude subscale (WS).

A pilot sample was collected with the HFS, followed by item analysis and item revisions using Nielsen and Kreiner's (2013) strategy for item modification; as a result, one item was excluded from the TCA, and a further three TCA items and two WS items were slightly adjusted. Thus, the resulting HFS-R consisted of 28 items measuring TCA, FAH, IA, and WS (see the above)

2.3. Participants and data collections

Our data were collected from university students in the health and social science faculties for both the pilot and the main study (Table 1). All data were collected by statistics teachers in their own classes. Students were informed about the purpose of the study and that participation was voluntary. The HFS was used for the pilot (N = 278) and the revised version, HFS-R, for the main study (N = 264). In both samples, high response rates (above 70%) were achieved, except for medical students in the pilot.

2.4. Rasch measurement models

The simplest model in the large family of IRT models is the Rasch model (RM) for dichotomous items (Rasch, 1960). In the present study, we used the partial credit model (PCM; Masters, 1982), which is a generalization of the Rasch model for ordinal data, as well as graphical log-linear Rasch models (Kreiner & Christensen, 2002, 2004, 2007), both as implemented in Digram (Kreiner, 2003; Kreiner & Nielsen, 2013). The Rasch model and the PCM generalization hold the same requirements for measurement (Kreiner, 2013; Mesbah & Kreiner, 2013); thus, we hereafter just use the term "RM" for Rasch model. The measurement requirements are as follows: i) unidimensionality—that the items of a scale measure only one underlying latent construct; ii) monotonicity—that the expected item score is a monotonically increasing function of the latent score; iii) local independence of items—that items are conditionally

Table 1. Data samples

	Pilot sample[a]		Revision sample[b]	
Academic program (semester)	**Possible N**	**Response rate % (total N)**	**Possible N**	**Response rate % (Total N)**
Medicine (2nd)	48	22.9		
Sociology (1st)	116	81.9	132	68.2
Sociology (3rd)			100	80.0
Master of health (*)	30	73.3		
Master of health (2nd)	30	70.0		
Public health (5th)	54	88.9	72	70.8
Public health (7th)			54	79.6
Total	278	70.9 (197)	358	73.7 (264)

Notes. *Preparatory course prior to program. a. Item analyses on the pilot data included 186 cases, as we decided to omit the medical students due to their low number. b. Item analyses on the revision sample were conducted with a slightly varying number of cases, as some had missing data; FAH 258 case, IA 240 cases, TCA 258 case, WS 252 cases.

independent given the latent score; iv) absence of DIF—that items and exogenous variables are conditionally independent given the latent score; and v) homogeneity—that the rank order of item parameters is the same for all persons no matter their level on the latent variable. The first four requirements are common for all IRT models, and they also define criterion-related construct validity (Rosenbaum, 1989). The fifth requirement is exclusive to the RM. Fulfillment of the five requirements by a set of item responses implies that the sum score is statistically sufficient for the person parameter (latent score), in the sense that all information on the person parameter is contained in the sum score over all items. Sufficiency of the sum score is a property only of the RM, as is the specifically objective nature of the set of item responses. Specific objectivity refers to the fact that within the frame of reference (hence the term "specific"), item comparisons do not depend on the persons, and person comparisons do not depend on the items (Rasch, 1961). The properties of RMs put together make the RM suitable as a means of data reduction when constructing simple and practical summated scales, such as the ones in this study (Nielsen, Kyvsgaard, Sildorf, Kreiner, & Svensson, 2017).

If fit to an RM is rejected, it is possible to achieve close to optimal measurement, provided that the departures from the RM consist exclusively of uniform DIF and/or uniform local dependence (uniform LD) between items (Kreiner & Christensen, 2007). Uniform/nonuniform refers to the way items depend either on exogenous variables or on other items. Thus, uniform implies that this dependence is the same across all levels of the latent variable, while nonuniform implies that it is not. If LD or DIF is uniform, the DIF or LD terms can be included and adjusted for in a so-called "Graphical Loglinear Rasch Model" (GLLRM), which is simply an extension of the RM allowing for exactly these two types of departures from the pure RM. A GLLRM adjusted only for uniform LD can retain sufficiency of the sum score, but the reliability of the scale will be negatively affected to some degree. A GLLRM adjusted for uniform DIF does not retain sufficiency of the sum score for the person parameter, as additional information on membership of subgroups for which items function differentially is also needed. However, this can be resolved by equating the sum score for the DIF discovered via the subgroup-dependent person parameters to allow unconfounded statistical comparisons (Kreiner, 2007).

2.4.1. Item analysis by RMs and GLLRMs

Item analyses of the four scales were all conducted using the same general strategy. First, the fit of the item responses of a given scale to the RM was tested. If fit to the RM was rejected, departures from the model were catalogued. If departures consisted of uniform LD and/or uniform DIF, then the fit of the item responses to a GLLRM adjusting for the discovered departures was tested. At a more detailed level, analysis included overall fit, DIF analysis at an overall and detailed

level, the effect of DIF on the sum score (in cases of DIF), analysis of local independence, item fit, unidimensionality, and analysis of reliability and targeting. Details of procedures and tests are provided below.

Global tests of fit (i.e. testing the equality of item parameters in low- and high-scoring groups) as well as global tests of no DIF were conducted using Andersen's (1973) conditional likelihood ratio (CLR) test. The fit of individual items was tested using conditional infit and outfit statistics (Christensen & Kreiner, 2013; Kreiner & Nielsen, 2013) and by comparing the observed item-rest-score correlations with the expected item-rest-score correlations under the model (Kreiner, 2011). The presence of DIF and LD in GLLRMs was also tested with Kelderman's (1984) likelihood-ratio test and partial Goodman–Kruskal gamma coefficients (Kreiner & Christensen, 2004). In the main study, DIF was tested specifically in relation to students' mathematics prerequisites (adequate, inadequate), academic discipline (sociology, public health), statistics course (first, second), age groups (19–20, 21–23, 24+), and gender (male, female). Unidimensionality was tested by comparing the expected correlation of subscales under the assumption that they measured the same underlying latent construct with the observed correlation (Horton, Marais, & Christensen, 2013), using a Monte Carlo approach for exact *p*-values.

In cases with fit to the RM, reliability was estimated as Cronbach's alpha, as this is known to provide the lower limit for reliability when items are locally independent. Accordingly, in cases with fit to a GLLRM, where items were not locally independent, we estimated reliability using Hamon and Mesbah's (2002) Monte Carlo method, as this takes into account any LD and adjusts the reliability accordingly. Targeting (i.e. the degree to which the study population was outside the target range) was assessed graphically as well as by two targeting indices. Item maps are plots of the distribution of person parameters and the distribution of item parameters onto the same latent scale. Thus, item maps allow visual evaluation of whether the majority of persons in the study population are included in the range of item parameters. Two numerical indices of the targeting of the person parameter were calculated, as described in Kreiner and Christensen (2013): the test information target index (i.e. the mean test information divided by the maximum test information) and the root mean squared error (RMSE) target index (i.e. the minimum standard error of measurement (SEM) divided by the mean SEM). Both indices should preferably be close to one. In addition, the target of the observed score and the SEM of the observed score were estimated.

Evidence of fit and no DIF discovered with the global tests was rejected if not supported by further evidence of item fit and lack of evidence of both DIF and LD in the detailed analysis. The Benjamini–Hochberg procedure was used to adjust for false discovery rate (FDR) due to multiple testing, whenever appropriate (Benjamini & Hochberg, 1995). Significance was evaluated at a 5% critical level (after adjusting for FDR). However, in line with the recommendations made by Cox et al. (1977), we distinguished among weak (*p*-values larger than .01), moderate (*p*-values larger than .001), and strong (*p*-values smaller than .001) evidence against the model.

2.4.2. Chain-graph models

In line with the suggestion by Kreiner and Christensen (2007), chain-graph models were used to illustrate the resulting models for each subscale. In these models, missing edges between variable nodes illustrate that the variables are conditionally independent, given the remaining variables in the model. Thus, two items that are not connected by an edge are conditionally independent of the remaining items given the latent variable (i.e. the property of no local dependence). Likewise, an item and a background variable that are not connected by an edge are conditionally independent given the latent variable and the other variables in the model (i.e. the property of no DIF). Undirected edges in the chain-graph models illustrate that the variables are conditionally dependent though with no causality assumed (e.g. in the case of locally dependent items). Directed edges (arrows) illustrate that variables are conditionally dependent while assuming causality (e.g.

in the case of DIF or of a background variable being related to the latent variable). Furthermore, when background variables are found to be associated directly with the latent variable, this can be seen as an assessment of criterion validity (i.e. that the latent scale score varies for a subgroup of respondents in an expected pattern).

2.5. Criterion validity

Concerning a priori expectations of associations between background variables and the latent scale score, we expected that students perceiving their level of mathematics ability to be inadequate for learning statistics would score higher on the anxiety scales and lower on the WS scale than would students perceiving that their level of mathematical ability is adequate. We had no a priori expectations concerning the relationship between the latent scale scores and the age and gender of the students, or with the academic discipline or whether it was the first or second statistics course.

To assess further the criterion validity of the final four scales measuring statistical anxiety and attitude toward statistics, Pearson correlations were calculated between all subscales for all students as well as for groups of students defined by their academic discipline and whether they were taking their first or second statistics course. Based on previous research (see the Introduction), we hypothesized 1) that the statistical anxiety subscales would be positively and significantly correlated and 2) that the statistical anxiety subscales and the attitude scale would be negatively and significantly correlated. Furthermore, 3) we expected to find these same patterns for all students and when stratifying the sample by academic discipline. Finally, we expected 4) that the negative correlations between the anxiety and attitude scales would be accentuated for students in their second statistics course. This we expected without assuming any causality between anxiety and attitude, as this might be in either direction or even be bidirectional. Chew and Dillon (2014) state that the general consensus in the field is that negative attitudes result in anxiety. Based on cognitive dissonance theory (Festinger, 1957), we suggest that the opposite relationship would be more probable, namely that long-term anxiety would increase negative attitudes toward the object of the anxiety.

3. Results

Only the results on the HFS-R from the main study are reported here, and as there are four subscales and thus many results, we have provided some of the results in a supplementary file.

The five-item FAH scale fits a pure RM with no evidence against homogeneity of the item parameters for high and low scorers, no evidence of DIF in relation to gender, age, academic discipline, prerequisite mathematical level, or whether the statistics course was the first or second one (Table 2). None of the TCA, IA, or WS scales fit pure RMs, as evidence of departures in the form of DIF and local dependence between item pairs was found. We thus proceeded with analysis by GLLRMs for these three subscales: the seven-item TCA scale, the eight-item IA scale, and a reduced six-item WS scale (one item was excluded to be used as a background variable and another was excluded during analysis due to lack of fit). Each fit a GLLRM with DIF and, in the case of the TCA and IA subscales, also with local dependence between items (Figure 1 and Table 2). The fit of individual items to the resulting models for each subscale is provided in Table S1 in the supplement.

3.1. Local dependence

The analyses revealed that four items in the TCA subscale were pairwise locally dependent, and that the partial γ correlations between these items were strong (Figure 1). The first locally dependent pair was item 1 (*Studying for an examination in a statistics course*) and item 3 (*Doing the final examination in a statistics course*) and the second pair was item 2 (*Doing the homework for a statistics course*) and item 4 (*participating in statistics exercises*). Analyses also identified two items in the IA subscale as locally dependent, but only with a medium strong partial γ correlation (Figure 1), namely item 2 (*Making an objective decision based on empirical data*) and item 5 (*To read and interpret output from a statistical analysis)*.

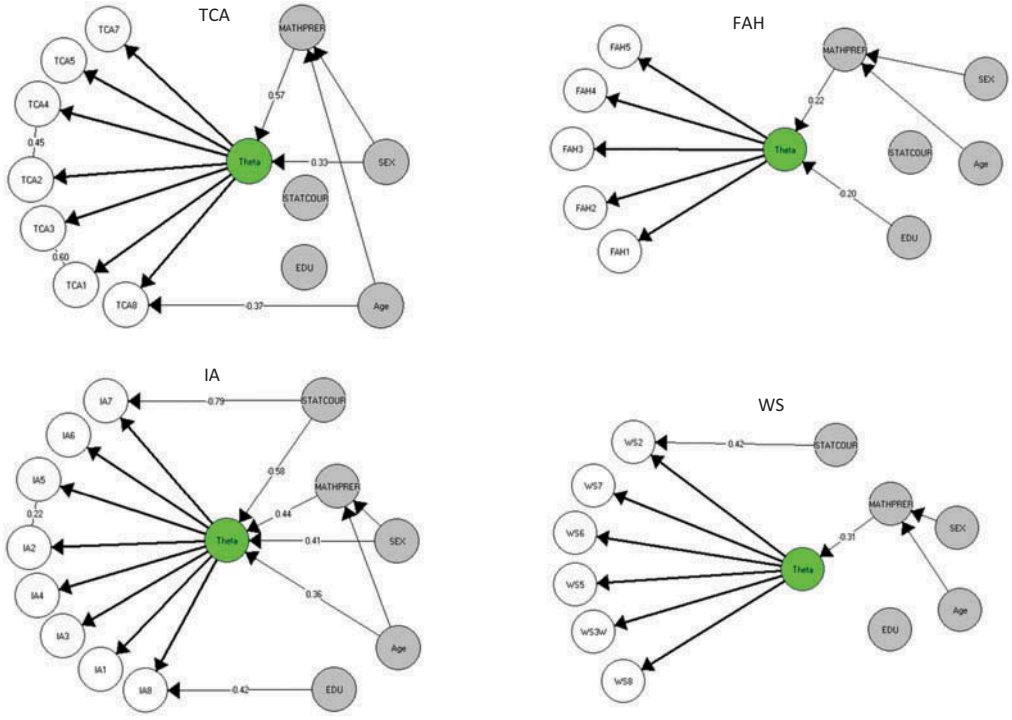

Figure 1. The final models for the four HFS-R subscales. γ-correlations are Goodman and Kruskal's rank correlation for ordinal data.

Table 2. The HFS-R: Global tests of fit and global tests of DIF for the final subscale models in Figure 1

	TCA GLLRM[a]			FAH RM			IA GLLRM[b]			WS GLLRM[c]		
Tests of fit	**CLR**	***Df***	***p***	**CLR**	***df***	***p***	**CLR**	***df***	***p***	**CLR**	***df***	***p***
Global homogeneity[+]	50.2	37	.072	22.8	12	.029	23.8	35	.924	16.4	20	.691
Global DIF												
Math prerequisite	43.6	37	.212	6.8	12	.868	44.1	35	.139	31.2	20	.053
Academic discipline	54.0	37	.035*	14.4	12	.279	31.3	30	.402	35.4	20	.018*
#Statistics course	47.0	37	.125	11.5	12	.491	36.4	29	.162	16.6	15	.343
Gender	43.4	37	.216	14.8	12	255	48.5	35	.064	32.9	20	.035*
Age	84.2	56	.009*	22.6	24	.541	60.8	70	.777	40.2	40	.461

[a]TCA model assuming that items 1 and 3, and 2 and 4, respectively, are locally dependent, and that item 8 functions differentially relative to age group.

[b]IA model assuming that items 2 and 5 are locally dependent, and that items 7 and 8 function differentially relative to academic discipline and statistics course, respectively.

[c]WS model assuming that item 2 functions differentially relative to statistics course.

[+]The test of homogeneity is a test of the hypothesis that item parameters are the same for persons with low or high scores.

*The Benjamini–Hochberg-adjusted critical level for DFR at the 5% level for the TCA and WS was .0083.

3.2. Effect of DIF

Evidence of DIF was discovered in the analysis of the TCA, IA, and WS subscales. Therefore, the raw scores of these subscales had to be equated for the DIF in order to eliminate the confounding effect of the background variables for which the DIF was discovered when using the summed scale score in subsequent statistical analysis. The score-equation tables are provided as Tables S2–S4 in the supplement, while the DIF results and the effect of adjusting for the DIF are presented in Tables 3–5.

3.2.1. The TCA subscale

In the TCA subscale, item 8 (*Going through an exam assignment in statistics after the grade has been given*) functioned differentially relative to age such that students were less likely to report high levels of anxiety in this situation the older they were, irrespective of their TCA level. Looking into the effect of equating the TCA raw score to adjust for the age DIF, it was clear that the difference in the mean scores for age groups became larger and more clearly significant as a result of the adjustment (Table 3). It was also evident that had there been just a few participants less, failure to adjust for the age DIF would have led to a type II error in a comparison of the TCA levels for students of different ages, as the (false) null hypothesis of no difference would have been accepted at a conventional 5% critical level.

Table 3. Comparison of observed and equated mean TCA scores in age groups

	Observed scores		Adjusted scores		
Age group (*n*)	**Mean**	**se**	**Mean**	**se**	**Bias**
19–20 years (43)	13.05	.46	13.05	.46	.00
21–23 years (136)	14.23	.29	14.40	.29	−.18
24 years and older (79)	14.41	.41	14.96	.38	−.55

Notes. Differences in observed mean scores (χ^2 (2) = 6.2, p = .045). Differences in adjusted mean scores (χ^2 (2) = 10.7, p = .005).

3.2.2. The IA subscale

Evidence of DIF relative to the study program was found for item 8 (*Seeing a student poring over the computer printouts from exercises*) in the IA scale, such that sociology students were systematically more likely to report high levels of anxiety in this situation than were the public health students, no matter their level of IA. Furthermore, evidence of DIF was also found for item 7 (*Figuring out whether to reject or retain the null hypothesis*) relative to statistics course, so that students in their first statistics course were systematically more likely to report high levels of anxiety in this situation than their second-course counterparts, regardless of their level of IA. Equating the IA raw scores to adjust for both instances of DIF relative to the statistics course did not affect the overall conclusion reached when comparing the observed mean scores and the adjusted mean scores for the four groups (Table 4). The comparison of observed mean scores and that of adjusted mean scores were significant. However, the bias resulting from the DIF was considered substantial, and thus DIF equating of the scores is necessary (Table 4).

Table 4. Comparison of observed and equated mean IA scores in groups defined by academic discipline and statistics course

	Observed scores		Adjusted scores		
Age group (*n*)	**Mean**	**se**	**Mean**	**se**	**Bias**
Sociology, 1st statistics course (72)	17.47	.40	17.47	.40	.00
Public health, 1st statistics course (49)	19.37	.39	19.76	.41	−.40
Sociology, not 1st statistics course (76)	15.39	.41	15.99	.42	−.59
Public health, not 1st statistics course (43)	17.09	.52	18.06	.55	−.97

Notes. Differences in observed mean scores four groups (χ^2 (3) = 49.9 p < .001). However, the first and last groups were not significantly different, and, if collapsed, the difference in observed means for the resulting three groups remained significant (χ^2 (2) = 49.6, p < .001). Differences in adjusted mean scores (χ^2 (3) = 42.3, p < .001). As with the observed means, the first and fourth groups could be collapsed as they were not significantly different, and the difference in adjusted scores in the three resulting groups also remained significant in this case (χ^2 (3) = 41.6, p < .001).

3.2.3. The WS scale

DIF relative to statistics course was discovered for item 2 (*I enjoy empirical work)* in the six-item WS scale, such that students in their second statistics course were systematically more likely to agree with this statement than were their first-course counterparts, independent of their level on the WS subscale. The effect of equating the WS raw score to adjust for the DIF relative to statistics course was modest, as the bias resulting from the DIF was also relatively small (Table 5). The comparison of the observed mean WS scores and that of the adjusted mean WS scores for students in their first and second statistics courses resulted in clearly significant differences.

Table 5. Comparison of observed and equated mean WS scores in statistics course groups

	Observed scores		Adjusted scores		
Statistics course group (*n*)	**Mean**	**se**	**Mean**	**se**	**Bias**
1st course (133)	21.50	.21	21.50	.21	.00
2nd course (119)	20.45	.26	20.21	.26	.23

Notes. Difference in observed mean scores (χ^2 (1) = 9.8, p = .002). Difference in adjusted mean scores (χ^2 (1) = 15.0, p < .001).

3.3. Targeting and reliability

Targeting was very good for the IA and TCA scales, although with some variations in subgroups defined by DIF variables. Thus, targeting of the TCA scales was best for students aged 24 years and older (84% of the maximum obtainable test information) and a little lower for the two younger groups of students (82% and 74% of the maximum obtainable test information, respectively) (Table 6). Item maps for the TCA scale in the three age groups demonstrate the very good targeting in different ways; Figure 2 panel A shows how the distributions of the person parameters and the item thresholds are better aligned in their location on the latent TCA scale for the two oldest subgroups in comparison to the youngest subgroup. Panel B of Figure 2 shows that even though the test information can be considered to be high, it is higher in the upper half of the estimated person parameter, when assuming that this is normally distributed. Finally, panel C in Figure 2 shows how the SEM is also rather high along the estimated person parameter under the assumption of a normal distribution, but highest at the low end of the latent TCA scale (the opposite of the test information). The variation in targeting of the IA scale was the best for public health students taking their first statistics course (95% of the maximum obtainable test information) and the poorest, although not actually bad, for sociology students taking their second statistics course (73%), with the two remaining groups falling in between these values (Table 6). The item maps for the IA scale show that the slightly worse targeting for sociology students taking their second statistics course was due to this subgroup being located more to the lower end of the latent IA scale, while most information was found at the higher end of the scale (Table S2, part 1, and part 2 in the supplement)

Targeting was poor for the FAH and WS scales, as only between 56% and 62% of the maximum obtainable test information was reached (Table 6). Students tended to be located in the lower end of the FAH theta scale compared to the location of the item thresholds, and for the WS scale the pattern was the reverse (Table S1 in supplement).

The reliability of the TCA scale was found to be at a less than satisfactory level (i.e. r < .70), as were the reliabilities for the IA scale in the groups of public health students taking their first statistics course and the reliability for the WS scale for students taking their first statistics course (Table 6). For the remaining subgroups in the IA and WS scales, and the FAH scale as a whole, reliability was satisfactory.

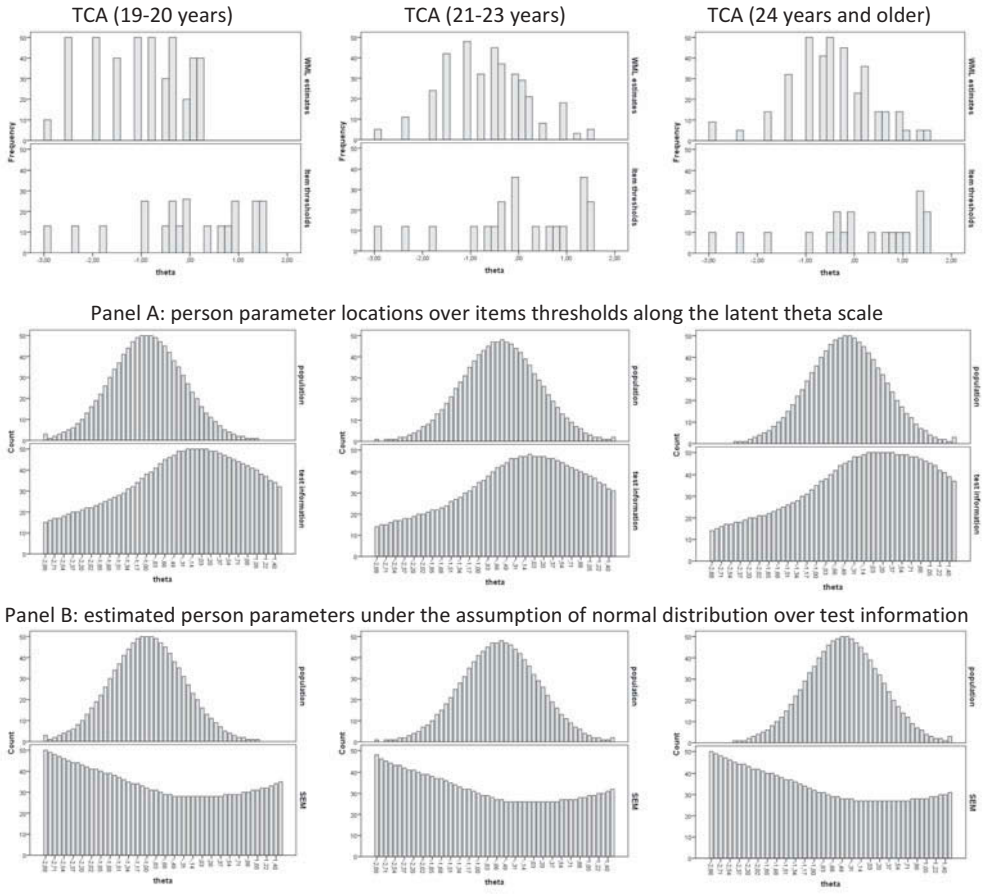

Figure 2. Item maps for the TCA scale in subgroups defined by age groups (DIF variable).

3.4. Unidimensionality of the anxiety subscales

To test the assumption of unidimensionality across the three anxiety subscales proposed in recent research (see the Introduction), pairwise tests of the hypothesis of unidimensionality were conducted for the three anxiety subscales. Unidimensionality was clearly rejected in all cases (Table 7).

3.5. Criterion validity

As expected, students perceiving their level of mathematics ability to be inadequate for learning statistics would score higher on the three statistical anxiety scales and lower on the WS scale than would students perceiving their level of mathematical ability to be adequate (Figure 1). Moreover, as expected, a priori and from the tests of unidimensionality (Table 7), the correlations between the three anxiety subscales were all positive and significant (Table 8). The highly significant correlations for all students did, however, present a more diverse picture when stratified by both academic discipline and statistics course. Thus, the FAH and TCA scales were more strongly correlated for public health students than for sociology students regardless of whether they were taking their first or second statistics course. The correlation between the IA and TCA scales was markedly higher for sociology students taking their second statistics course than for the remaining three groups of students. The correlation between the IA and FAH scales did not vary much across subgroups of students.

The correlations between the three anxiety subscales (TCA, FAH, and IA) on one side and the attitude scale (WS) on the other side partially confirmed our hypothesis that these would be negatively correlated (Table 8). Thus, for the entire groups of students, this was indeed the case, although only weak negative correlations were found, and only significant in two cases. However,

Table 6. Targeting and reliability of the TCA, FAH, IA, and WS scales

	Theta								Sum score			
Groups defined by DIF[a]	**Target**	**Mean**	**TI mean**	**TI max**	**TI Target index**	**RMSE mean**	**RMSE min**	**RMSE target index**	**Target**	**Mean**	**Mean SEM**	**r**
Test and Class Anxiety (TCA)												
19–20 years	−.08	−.98	3.401	4.599	.739	.561	.466	.831	16.47	13.05	1.83	.64
21–23 years	−.05	−.63	3.740	4.612	.811	.535	.466	.871	16.39	14.23	1.92	.67
24 years and older	−15	−.46	3.764	4.474	.841	.531	.473	.890	16.76	14.41	1.93	.64
Fear of Asking for Help (FAH)												
All *(no DIF)*	.27	−1.69	1.618	2.748	.589	.781	.603	.772	12.05	8.09	1.22	.81
Interpretation Anxiety (IA)												
Sociology, 1st statistics course	.69	−.52	3.136	3.685	.851	.573	.521	.910	21.56	17.47	1.77	.75
Public health, 1st statistics course	.68	.20	3.374	3.557	.949	.547	.530	.970	21.06	18.37	1.84	.56
Sociology, 2nd statistics course	1.35	−1.04	2.785	3.836	.726	.618	.511	.826	23.33	15.39	1.66	.79
Public health, 2nd statistics course	1.23	−.34	3.049	3.670	.831	.582	.522	.897	22.31	17.09	1.92	.75
Worth of Statistics (WS)												
1st statistics course	−.87	1.64	1.956	3.483	.562	.738	.536	.726	12.91	19.50	1.38	.66
2nd statistics course	−1.02	1.05	2.369	3.811	.622	.676	.676	.758	12.39	18.45	1.52	.71

Notes. TI, test information; RMSE, the root mean squared error of the estimated theta score; SEM, standard error of measurement of the observed score; R, reliability

[a]. For the TCA, IA, and WS scales, targeting and reliability are provided for groups defined by DIF variables.

cogent education

Table 7. Test of unidimensionality of anxiety subscales

Anxiety Subscales	Observed γ	Expected γ	SE expected γ	Asymptotic *p*	Exact *p*
TCA & IA	.426	.533	.039	<.01	<.001
TCA & FAH	.324	.491	.043	<.001	<.001
IA & FAH	.505	.193	.043	<.0001	<. 001

Notes. γ, gamma correlation between subscales; observed and expected under the model. γ-correlations are Goodman and Kruskal's rank correlation for ordinal data. * parametric bootstrapping with 400 samples.

Table 8. Pearson correlations for all students and stratified by academic discipline and statistics course

Scales and groups of students	TCA	FAH	IA
FAH			
all students	.445***		
sociology, 1st statistics course	.385***		
public health, 1st statistics course	.607***		
sociology, 2nd statistics course	.442***		
public health, 2nd statistics course	.533***		
IA all students	.547***	.279***	
sociology, 1st statistics course	.530***	.346**	
public health, 1st statistics course	.541***	.381**	
sociology, 2nd statistics course	.702***	.308**	
public health, 2nd statistics course	.423**	.284*	
WS			
all students	−.167**	−.054	−.136*
sociology, 1st statistics course	.005	.059	−.048
public health, 1st statistics course	−.118	−.159	−.111
sociology, 2nd statistics course	−.327**	−.103	−.403***
public health, 2nd statistics course	−.308*	−.158	−.391**

Notes. * $p < .05$, ** $p < .01$, *** $p < .001$, one-tailed *p*-values.

the stratification by both academic discipline and statistics course revealed an interesting pattern: for all groups of students, the correlations between WS and FAH were very low and not significant. With regard to the correlations between WS and TCA, these were weak and insignificant for students taking their first statistics course, but moderate, negative, and significant for students in their second statistics course, no matter their academic discipline. Similarly, the correlations between WS and IA were weak and nonsignificant for students taking their first statistics course, but moderate, negative, and significant for students in their second statistics course, again regardless of their academic discipline. Thus, the expected accentuation of the negative relationship between attitudes toward statistics and statistical anxiety for students in their second statistics course was indeed present, although only for TCA and IA.

4. Discussion and implications for further validity studies

The four subscales of the HFS-R were each found to fit either a pure RM or a GLLRM, with only a few instances of DIF and violations of local independence between items (LD). As all instances of DIF

and LD were uniform in nature, these could easily be adjusted for to make both the latent scales scores (i.e. person parameters) and the observed scale scores (i.e. the summed raw scores) comparable across subgroups. Furthermore, it was rejected that the three anxiety subscales (TCA, FAH, and IA) made up one overall and unidimensional statistical anxiety subscale. A priori hypotheses on subscale correlations for different subgroups of students were partly supported. Taken together, the findings supported both construct and criterion validities of the four-subscale HFS-R instrument. Furthermore, the reliability of the subscales was satisfactory for two subscales and was overall comparable to previous findings. Accordingly, we found the four subscales of the HFS-R to be of a sufficient psychometric quality as to warrant their use for evaluating statistical anxiety and attitude toward statistics at the class level as an educational tool for statistics teachers of higher education students in other disciplines than statistics. When used for this purpose in a single statistics class, only the TCA scale appears in need of adjustment for DIF relative to age, as results at the class level will otherwise appear somewhat underestimated due to the artificially low scores of the older students. Additionally, we found that the HFS-R was suitable for research on higher education students' attitudes and relationship to statistics, when appropriately adjusted for DIF. Future research should explore the value of statistics teachers obtaining knowledge of the different types of statistical anxiety as well as the attitude toward statistics of their students and the possible effects of antianxiety interventions based on this knowledge. As the HFS-R has been developed and validated in a Danish language version, but the instrument is also available in an English version, we recommend that validity studies should be conducted and that intervention studies should also be undertaken outside the Danish university context. A further and natural extension would also consist of translations into other languages (e.g. the other Scandinavian languages) and subsequent cross-cultural validity studies using IRT.

The HFS-R consists of four subscales measuring TCA (seven items), FAH (five items), IA (eigth items), and WS (six items), totaling 26 items, and with only four items suffering from DIF (one relative to age, one relative to academic discipline, and two relative to whether the statistics course was the first or the second taken). This is a substantially shorter questionnaire than the original STARS (51 items), and not only because the STARs included two additional attitude subscales. In comparison to the one validity study of the STARS employing a thorough IRT approach (i.e. Teman, 2013), the HFS-R as a whole and subscale-wise includes more items. Teman's study, after exclusion of non-fitting items, resulted in a 31-item instrument, where six items suffered from DIF relative to study level (undergraduate versus graduate). In comparison to the present study, the TCA, FAH, IA, and WS subscales in Teman's study totaled just 21 items, with three of the items suffering from DIF. Looking further into the specific items excluded in the present study and Teman's study, there is agreement on the exclusion of 10 items across the four subscales (items 9, 10, 12, 19, 24, 36, 42, 47, 49, and 50), and a further three of the items excluded by Teman were items that were modified in the present study (items 18, 29, and 33). This is a substantial degree of agreement when taking into account that the approaches, strategies, and IRT models differed in the two studies. For some items the agreement of exclusion might be due to the fact that they have become outdated—e.g. in the case of item 12, *Arranging to have a body of data put into the computer,* and item 19, *Asking someone in the computer lab for help in understanding a printout* (all item texts can be seen in Teman, 2013). Such obsolescence was indeed a main cause for exclusion in the present study, and should also result in non-fit of the items, as they will not appear to be measuring the constructs in question in a contemporary manner.

In the present study, we found strong evidence against unidimensionality across the three anxiety subscales: TCA, FAH, and IA. Therefore, it is not appropriate to combine these three subscales of the HFS-R into one composite scale, as suggested by Hsiao (2010) for a Chinese translation of the STARS, and as subsequently proposed by Papousek and colleagues (Macher et al., 2011, 2013; Papousek et al., 2012) for a German translation of the STARS. The present results are, of course, not directly corresponding to the studies by Hsiao or Papousek and colleagues, as we employ not only a third translation of the STARS, but also an adapted version. We cannot with certainty say that our rejection of unidimensionality of the three anxiety scales would extend to

the original STARS subscales, as we have not only translated but also modified some items. We do, however, propose that such an explicit testing of the assumption of unidimensionality across anxiety (and attitude) subscales of the STARS should be undertaken in future research. Furthermore, we suggest that unidimensionality should be assessed after first having ascertained whether items of the single subscales actually fit the scales and whether items suffer from DIF, and as the only other IRT study of the STARS (Teman, 2013) has proposed that many items in the original STARS do not fit the subscale models or suffer from DIF.

With regard to targeting, we found that both the IA and TCA scales were well targeted to the study population, while the FAH and WS scales were not. These findings do not reflect that the FAH and WS scales are of poorer psychometric quality than were the IA and TCA scales. Rather, they reflect that both the sociology and public health students engaged in their first or second statistics course score in parts of these scales where the test information is less than optimal. Thus, students are located more toward the lower end of the FAH scale, whereas there is most information at the higher end of this scale. On the WS scale, students were located more to the center of the scale, while the most information was at the lower end of this scale. In other words, had the students in the study sample exhibited more FAH and had they had a less positive opinion of the worth that can be attributed to statistics, then these scales would have been better targeted for these particular student groups. Thus, we recommend that future validity studies should attempt to include students from academic disciplines where a higher degree of FAH would be expected and/or where a lesser value is placed on statistics in the student discourse.

The reliability of the TCA scale was found to be less than satisfactory using conventional cut-points, as was the reliability of the IA scale for public health students taking their first statistics course, and the reliability of the WS scale for students taking their first statistics course. All reliabilities in the present study, except for the FAH scale, were found to be lower than those reported in the original study by Cruise et al. (1985), who obtained reliabilities of .91 (TCA), .89 (IA), .85 (FAH), and .94 (WS). More recent studies utilizing the original English version of the STARS (Teman, 2013), a Turkish, a Chinese, and a German translation (Baloglu, Abbassi, & Kesici, 2017; Hsiao, 2010; Papousek et al., 2012), reported reliabilities of the TCA scale ranging from .81 to .87, of the IA scale ranging from .83 to .87, and of the WS scale ranging from .82 to .94. As these studies all utilized slightly longer TCA and IA scales and a substantially longer WS scale than in the present study, the differences are not surprising. The reliability of the FAH scale was reported to be in the range of .77–.86 in the above-mentioned studies, which is comparable to the present finding, where we used a five-item modified FAH scale as opposed to the original four-item FAH scale.

The modest criterion validity study included here did, as expected, show that the three anxiety scales were positively and significantly correlated. This was found both in connection with the unidimensionality test, where observed subscale correlations in the form of gamma rank correlations were compared to the expected subscale correlations under the model, and in the subsequent correlational analyses using Pearson's product-moment coefficients. Recent studies, using different language versions of the STARS (i.e. English, German, and Turkish), with differing samples of university students, have also reported positive correlations between the three anxiety scales, although of somewhat varying magnitudes (Baloglu et al., 2017; DeVaney, 2016; Papousek et al., 2012).

Additionally, and more interestingly, the criterion validity part of the present study also showed that the correlational relationship between the three anxiety subscales and the attitude subscale (i.e. WS) was as expected in the cases of the TCA and IA scales, but not in the case of the FAH scale, which was not correlated with WS at all. Furthermore, it was confirmed that the negative correlation between anxiety and attitude was accentuated for students engaged in their second statistics course, as there were only weak and insignificant correlations for students in their first statistics course. This finding might cautiously be suggested to arise from a cognitive dissonance

phenomenon (Festinger, 1957), in the sense that prolonged anxiety will lead to an appropriate adjustment of attitude. We want to stress that this cannot be concluded from the present study as we do not have longitudinal data and therefore cannot infer causality between anxiety and attitude. However, the same relationship between the WS and the three attitude scales has been reported in recent studies as positive correlations resulting from the original WS scale containing negatively worded items rather than positively worded items as in the present study (Baloglu et al., 2017; DeVaney, 2016; Papousek et al., 2012). Thus, we suggest that future studies should include longitudinal measurements, and preferably at more than two time-points, in order to look further into any time-wise causal relationship between statistical anxiety and attitudes toward statistics. Such studies might also include qualitative elements (e.g. interviews and/or observations).

Supplementary material
Supplemental data for this article can be accessed here.

Acknowledgements
We would like to thank Robert J. Cruise for providing us with the unpublished STARS questionnaire and allowing us to work with it in our study. We are grateful to Cecilie Kreiner for her translation of the STARS into Danish and her efforts in providing us with English version of the new items, which we have developed as part of the study. We also acknowledge the (other) statistics teachers, who were kind enough to collect data for us, and repeatedly so: Bella Marckmann, Lars Pico, Karl Bang Christensen, and Theis Lange.

Funding
The author received no direct funding for this research.

Author details
Tine Nielsen[1]
E-mail: tine.nielsen@psy.ku.dk
Svend Kreiner[2]
E-mail: svkr@sund.ku.dk
[1] Department of Psychology, University of Copenhagen, Copenhagen, Denmark.
[2] Biostatistics Unit, Department of Public health, University of Copenhagen, Copenhagen, Denmark.

The HFS-R is available in both a Danish and an English version from the corresponding author.

Citation information
Cite this article as: Measuring statistical anxiety and attitudes toward statistics: The development of a comprehensive Danish instrument (HFS-R), Tine Nielsen & Svend Kreiner, *Cogent Education* (2018), 5: 1521574.

References

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*(1), 123–140. doi:10.1007/BF02291180
Andrich, D. (1978). A rating scale formulation for ordered response categories. *Psychometrika*, *43*, 561–573. doi:10.1007/bf02293814
Baloglu, M., Abbassi, A., & Kesici, S. (2017). Multivariate relationships between statistics anxiety and motivational beliefs. *Education*, *137*(4), 430–444.
Baloğlu, M., Deniz, M. E., & Kesici, S. (2011). A descriptive study of individual and cross-cultural differences in statistics anxiety. *Learning and Individual Differences*, *21*(4), 387–391. doi:10.1016/j.lindif.2011.03.003
Bendig, A. W., & Hughes, J. B. (1954). Student attitude and achievement in a course in introductory statistics. *Journal of Educational Psychology*, *45*, 268–276. doi:10.1037/h0057391
Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300.
Bui, N. H., & Alfaro, M. A. (2011). Statistics anxiety and science attitudes: Age, gender, and ethnicity factors. *College Student Journal*, *45*(3), 573–586.
Chew, P. K., & Dillon, D. B. (2014). Statistics anxiety update: Refining the construct and recommendations for a new research agenda. *Perspectives on Psychological Science*, *9*(2), 196–208. doi:10.1177/1745691613518077
Christensen, K. B., & Kreiner, S. (2013). Item fit statistics. In K. B. Christensen, S. Kreiner, & M. Mesbah (Eds.), *Rasch Models Health* (pp. 83–104). London: ISTE and John Wiley & Sons Inc. doi:10.1002/9781118574454.ch5
Cox, D. R., Spjøtvoll, E., Johansen, S., van Zwet, W. R., Bithell, J. F., & Barndorff-Nielsen, O., et al. (1977). The role of significance tests [with discussion and reply]. *Scandinavian Journal of Statistics,4*(2), 49–70.
Cruise, R. J., Cash, R. W., & Bolton, D. L. (1985). Development and validation of an instrument to measure statistical anxiety. *Paper Presented at the Proceedings of the American Statistical Association*.
Cruise, R. J., & Wilkins, E. M. (1980) *STARS: Statistical anxiety rating scale*, Unpublished manuscript, Andrews University, Berrien Springs, MI.
DeVaney, T. A. (2016). Confirmatory factor analysis of the statistical anxiety rating scale with online graduate students. *Psychological Reports*, *118*(2), 565–586. doi:10.1177/0033294116644093
Earp, M. S. (2007). *Development and validation of the statistics anxiety measure* (Unpublished doctoral dissertation). University of Denver, CO.
Festinger, L. (1957). A theory of cognitive dissonance. Stanford, California: Stanford University Press.
Galli, S., Ciancaleoni, M., Chiesi, F., & Primi, C. (2008). Who failed the introductory statistics examination? A study on a sample of psychology students. Paper presented at the 11th International Congress on Mathematical Education, Monterrey, Mexico.
Hamon, A., & Mesbah, M. (2002). Questionnaire reliability under the rasch model. In M. Mesbah, B. F. Cole, & M. L. T. Lee (Eds.), *Statistical methods for quality of life studies. Design, measurement and analysis* (pp. 155-168). Dordrecht: Kluwer Academic Publishers. doi:10.1007/978-1-4757-3625-0_13
Horton, M., Marais, I., & Christensen, K. B. (2013). Dimensionality. In K. B. Christensen, S. Kreiner, & M. Mesbah (Eds.), *Rasch Models Health* (pp. 137–158). London: ISTE and John Wiley & Sons Inc., doi:10.1002/9781118574454.ch9
Hsiao, T. Y. (2010). The statistical anxiety rating scale: Further evidence for multidimensionality. *Psychological Reports*, *107*(3), 977–982, doi:10.2466/07.11.pr0.107.6.977-982
Hsiao, T. Y., & Chiang, S. (2011). Gender differences in statistics anxiety among graduate students learning

English as a foreign language. *Social Behavior and Personality: an International Journal*, *39*(1), 41–43. doi:10.2224/sbp.2011.39.1.41

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika*, *49*, 223–45. doi:10.1007/BF02294174.

Kesici, Ş., Baloğlu, M., & Deniz, M. E. (2011). Self-regulated learning strategies in relation with statistics anxiety. *Learning and Individual Differences*, *21*(4), 472–477. doi:10.1016/j.lindif.2011.02.006

Kreiner, S., & Christensen, K. B. (2002). Graphical Rasch models. In M. Mesbah, B. F. Cole, & M. T. Lee (Eds.), *Statistical methods for quality of life studies* (pp. 187–203). Dordrecht: Kluwer Academic Publishers. doi:10.1007/978-1-4757-3625-0_15

Kreiner, S. (2003). *Introduction to DIGRAM. Research report 03/10. Copenhagen, department of biostatistics*. Copenhagen: University of Copenhagen.

Kreiner, S., & Christensen, K. B. (2007). Validity and objectivity in health-related scales: Analysis by graphical loglinear rasch models. In von Davier & Carstensen (Eds.), *Multivariate and mixture distribution rasch models* (pp. 329–346). New York: Springer. doi:10.1007/978-0-387-49839-3_21

Kreiner, S. (2007). Validity and objectivity: Reflections on the role and nature of Rasch models. *Nordic Psychology*, *59*(3), 268–298. doi:10.1027/1901-2276.59.3.268

Kreiner, S. (2011). A note on item-restscore association in rasch models. *Applied Psychological Measurement*, *35*(7), 557–561. doi:10.1177/0146621611410227

Kreiner, S., & Christensen, K. B. (2013). Person parameter estimation and measurement in Rasch models. In K. B. Christensen, S. Kreiner, & M. Mesbah (Eds.), *Rasch Models Health* (pp. 63–78). London: ISTE and John Wiley & Sons Inc. doi:10.1002/9781118574454.ch4

Kreiner, S. (2013). The rasch model for dichotomous items. In K. B. Christensen, S. Kreiner, & M. Mesbah (Eds.), *Rasch Models Health* (pp. 5–26). London: ISTE and John Wiley & Sons Inc. doi:10.1002/9781118574454.ch1

Kreiner, S., & Christensen, K. B. (2004). Analysis of local dependence and multidimensionality in graphical loglinear Rasch models. *Communication in Statistics – Theory and Methods*, *33*(6), 1239–1276. doi:10.1081/sta-120030148

Kreiner, S., & Nielsen, T. (2013). *Item analysis in DIGRAM 3.04. Part I: Guided tours. Research report 2013/06. Department of public health*. Copenhagen: University of Copenhagen.

Maat, S. M., & Rosli, M. K. (2016). The Rasch model analysis for statistical anxiety rating scale (STARS). *Creative Education*, *7*, 2820–2828. doi:10.4236/ce.2016.718261

Macher, D., Paechter, M., Papousek, I., & Ruggeri, K. (2011). Statistics anxiety, trait anxiety, learning behavior, and academic performance. *European Journal of Psychology of Education*, 1–16. ISSN 0256–2928. doi:10.1007/s10212-011-0090-5

Macher, D., Paechter, M., Papousek, I., Ruggeri, K., Freudenthaler, H. H., & Arendasy, M. (2013). Statistics anxiety, state anxiety during an examination, and academic achievement. *British Journal of Educational Psychology*, *83*(4), 535–549. doi:10.1111/j.2044-8279.2012.02081.x

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174. doi:10.1007/bf02296272

Mesbah, M., & Kreiner, S. (2013). The Rasch model for ordered polytomous items. In K. B. Christensen, S. Kreiner, & M. Mesbah (Eds.), *Rasch Models Health* (pp. 27–42). London: ISTE and John Wiley & Sons Inc. doi:10.1002/9781118574454.ch2

Nielsen, J. B., Kyvsgaard, J. N., Sildorf, S. M., Kreiner, S., & Svensson, J. (2017). Item analysis using Rasch models confirms that the Danish versions of the DISABKIDS® chronic-generic and diabetes-specific modules are valid and reliable. *Health and Quality of Life Outcomes*, *15*(1), 44. doi:10.1186/s12955-017-0618-8

Nielsen, T., & Kreiner, S. (2013). Improving items that do not fit the Rasch model: Exemplified with the physical functioning scale of the SF-36. Annales de L'I.S.U.P. *Publications De L'Institut De Statistique De L'Université De Paris, Numero Special*, *57*(1–2), 91–108.

Onwuegbuzie, A. J., & Seaman, M. A. (1995). The effect of time constraints and statistics test anxiety on test performance in a statistics course. *Journal of Experimental Education*, *62*, 115–124. doi:10.1080/00220973.1995.9943816

Papousek, I., Ruggeri, K., Macher, D., Paechter, M., Heene, M., Weiss, E. M., … Freudenthaler, H. (2012). Psychometric evaluation and experimental validation of the statistics anxiety rating scale. *Journal of Personality Assessment*, *94*(1), 82–91. doi:10.1080/00223891.2011.627959

Pretorius, T. B., & Norman, A. M. (1992). Psychometric data on the statistics anxiety scale for a sample of South African students. *Educational and Psychological Measurement*, *52*, 933–937. doi:10.1177/0013164492052004015

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. doi:10.1016/s0019-9958(61)80061-2

Rasch, G. (1961). *On General Laws and the Meaning of Measurement in Psychology*. The Regents of the University of California. Available from: http://projecteuclid.org/euclid.bsmsp/1200512895. Downloaded 06 June 2018.

Rosenbaum, P. (1989). Criterion-related construct validity. *Psychometrika*, *54*, 625–633. doi:10.1007/bf02296400

Teman, E. D. (2013). A Rasch analysis of the statistical anxiety rating scale. *Journal of Applied Measurement*, *14*, 4.

Vigil-Colet, A., Lorenzo-Seva, U., & Condon, L. (2008). Development and validation of the statistical anxiety scale. *Psicothema*, *20*(1), 174–180.

Wiiliams, A. S. (2010). Statistics anxiety and instructor immediacy. *Journal of Statistics Education*, *18*(2), 1–18. doi:10.1080/10691898.2010.11889795

Zeidner, M. (1991). Statistics and mathematics anxiety in social science students: Some interesting parallels. *British Journal of Educational Psychology*, *61*, 319–328. doi:10.1111/j.2044-8279.1991.tb00989.x

Cogent Education **(ISSN: 2331-186X) is published by Cogent OA, part of Taylor & Francis Group.**